كلية الحاسبات والذكاء الإصطناعي

BS111
Probability and Statistics

Lecture 13

Dr. Ahmed Hagag

Faculty of Computers and Artificial Intelligence
Benha University

2021

Ch 4: Descriptive Statistics

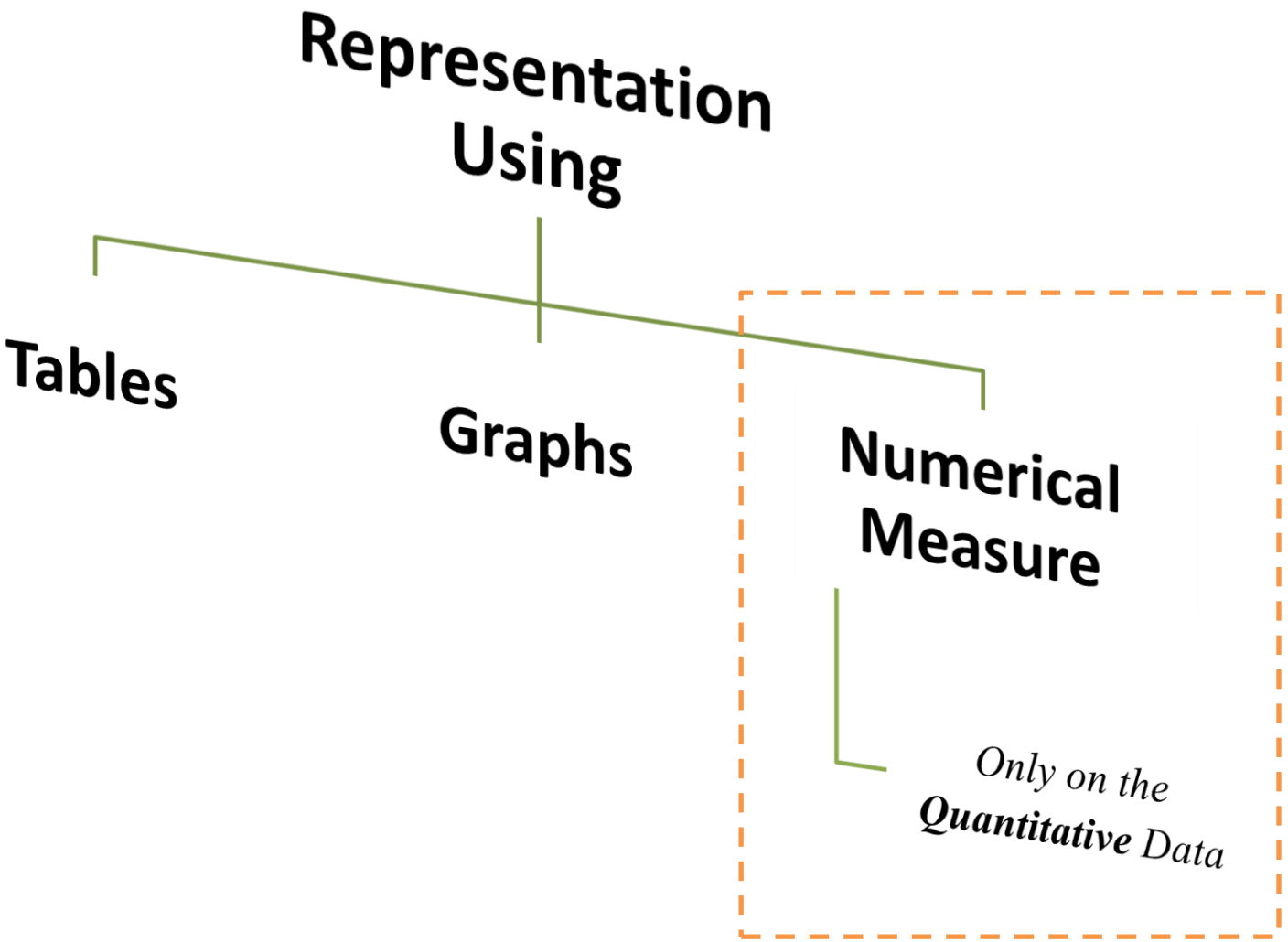
Data Description (1/2)

- Population and Sample in a Statistical Study.
- Types of Data.
- Representation of Data.
- Numerical Summaries of Quantitative Data.
  - Measures of Centrality (Central Tendency/location)
  - Measures of Variability (Dispersion/Spread)
  - Measures of Shape. (Skewness, Kurtosis)
  - Measures of Relative Position. (Percentiles, Quartiles, IQR, Coefficient of Variation).

Data Description (2/2)

- Stem-and-Leaf Diagrams.
- Box-Whisker Plot (Box Plot).
- Time Sequence Plots.
- Scatter Diagrams.

Recall


Representation
Using
Tables
Graphs
Numerical
Measure
Only on the
Quantitative Data

Numerical Summaries (1/3)

Measures of Centrality (1/2):

(Central Tendency / location)

- Measures of location are designed to provide the analyst with some quantitative values of where the center, or some other location, of data is located.
- We need **one** value to represent the data.

Numerical Summaries (1/3)

Measures of Centrality (2/2):

The following measures are of primary importance:

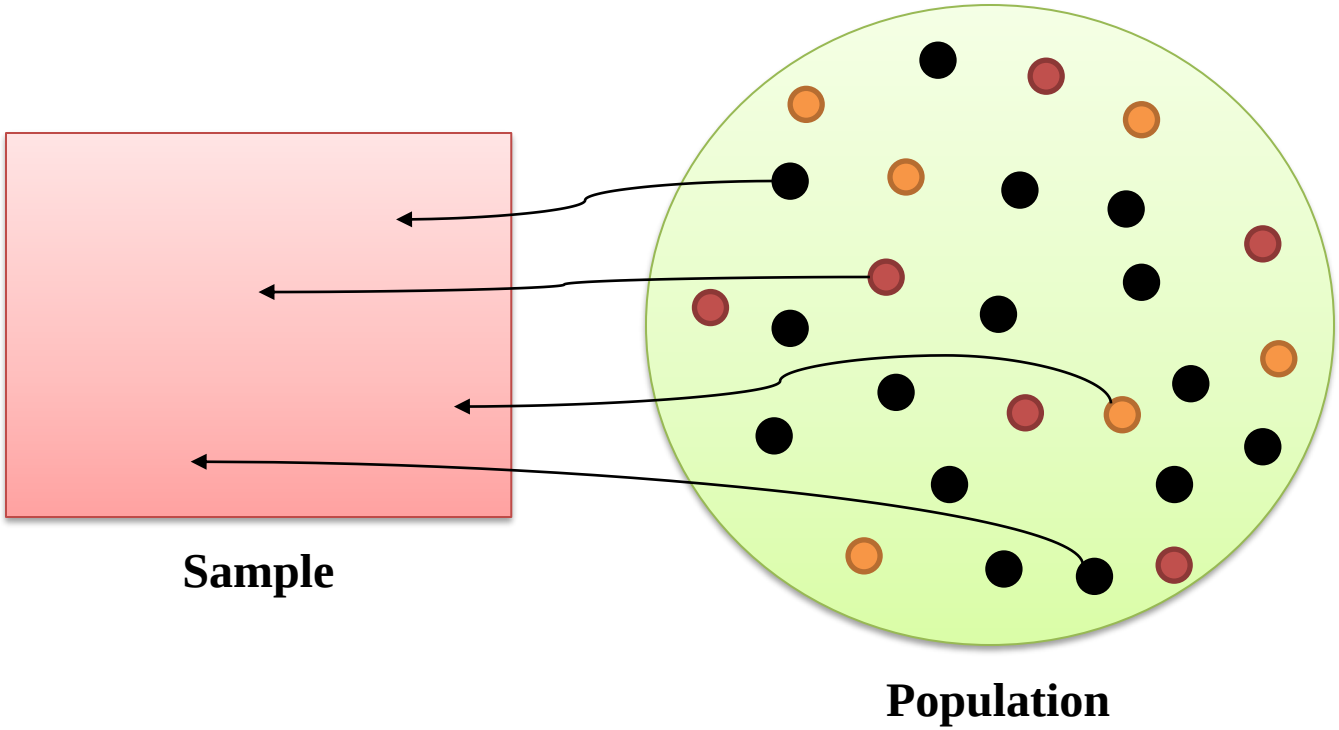
1. Mean
2. Median
3. Mode

The ***mean***, also sometimes referred to as the arithmetic mean, is the most useful and most commonly used measure of centrality. The ***median*** is the second most used, and the ***mode*** is the least used measure of centrality.

Numerical Summaries (2/3)

Measures of Centrality – Mean (1/9):

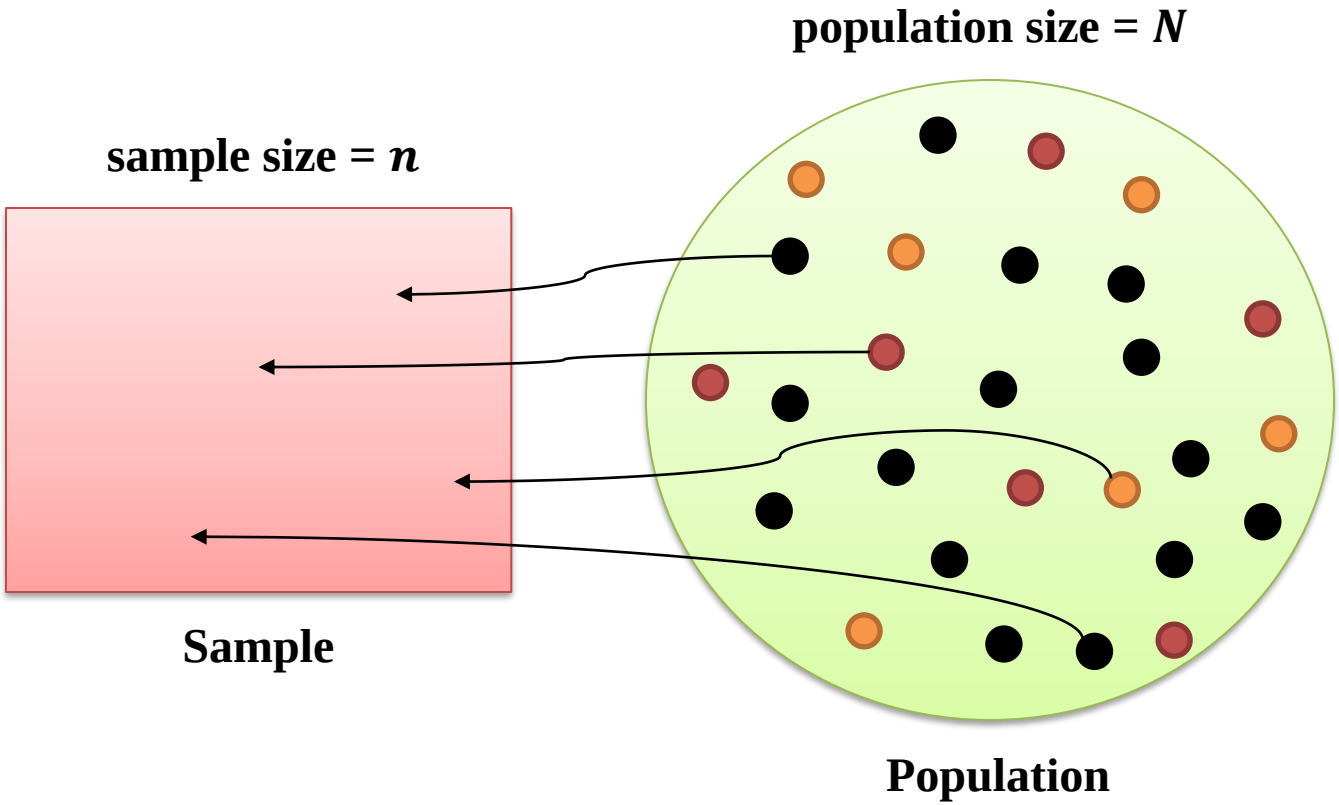
Sample or Population



Numerical Summaries (2/3)

Measures of Centrality – Mean (1/9):

Sample or Population



Numerical Summaries (2/3)

Measures of Centrality – Mean (2/9):

population mean

The population mean is denoted by the Greek letter μ (read as meu), for a finite population with N equally likely values:

$$\mu = \sum_{i=1}^{N} x_i f(x_i) = \frac{\sum_{i=1}^{N} x_i}{N}$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (3/9):

sample mean "sample average"

The **sample mean** is the **average** value of all observations in the data set. Usually, these data are a sample of observations that have been selected from some larger population of observations.

It is denoted by $\bar{x}$ (read as x bar), for n observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (4/9):

sample mean "sample average"

The sample mean is an **estimate** of the population mean.

the ***degrees of freedom*** associated with the mean estimate

It is denoted by $\bar{x}$ (read as x bar), for n observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (5/9):

Example 1:

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

The sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8}$$

$$= \frac{12.6 + 12.9 + \cdots + 13.1}{8} = \frac{104}{8} = 13.0$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (5/9):

Example 1:

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

The sample mean is:

Numerical Summaries (2/3)

Measures of Centrality – Mean (6/9):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Find the mean sale of these individuals.

Numerical Summaries (2/3)

Measures of Centrality – Mean (6/9):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

The mean sale of these individuals is:

$$\bar{x} = \frac{(10 + 8 + 15 + 12 + 17 + 7 + 20 + 19 + 22 + 25 + 16 + 15 + 18 + 250 + 300 + 12)}{16}$$

$$= 47.875$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (6/9):

Mean is sensitive to extreme values (**outliers**)

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

The mean sale of these individuals is:

$\bar{x} = 47.875$

In this case, the mean does not adequately represent the measure of centrality of the data set.

Numerical Summaries (2/3)

Measures of Centrality – Mean (7/9):

trimmed mean

A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values. We will denote it by $\bar{x}_{t=0.1}$ or $\bar{x}_{tr(10)}$

Numerical Summaries (2/3)

Measures of Centrality – Mean (8/9):

Example 1:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Find the 13% trimmed mean sale of these individuals.

Numerical Summaries (2/3)

Measures of Centrality – Mean (8/9):

Example 1:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Find the 13% trimmed mean sale of these individuals.

order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Remove 16*13/100 = 2.08 from each side.

Numerical Summaries (2/3)

Measures of Centrality – Mean (8/9):

Example 1:

order		10	12	12	15	15	16	17	18	19	20	22	25	

$$\bar{x}_{tr(13)} = \bar{x}_{t=0.13} = 16.75$$

$$\frac{(10 + 12 + 12 + 15 + 15 + 16 + 17 + 18 + 19 + 20 + 22 + 25)}{12}$$

Numerical Summaries (2/3)

Measures of Centrality – Mean (9/9):

weighted mean

Sometimes, we are interested in finding the sample average of a data set where each observation is given a relative importance expressed numerically by a set of values called weights.

$$\bar{X}_w = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_n X_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Numerical Summaries (2/3)

Measures of Centrality – Median (1/4):

Median (M)

The **median** ***M*** is the midpoint of a distribution. Half the observations are smaller than the median, and the other half are larger than the median.

Numerical Summaries (2/3)

Measures of Centrality – Median (2/4):

Median (M)

Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. Find the location (rank) of the median of a data set of size n

$$\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases}$$

3. Find the value of the observation corresponding to the rank of the median founded in step 2. i. e., $x_{(n+1)/2}$ or $(x_{n/2} + x_{n/2+1})/2$

Numerical Summaries (2/3)

Measures of Centrality – Median (3/4):

Example 1:

The nine observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, $x_8 = 13.1$ and $x_9 = 13.2$.

Arrange all observations in order of size $n = 9$:

Order	**12.3**	**12.6**	**12.6**	**12.9**	**13.1**	**13.2**	**13.4**	**13.5**	**13.6**
Ranks	1	2	3	4	5	6	7	8	9

Numerical Summaries (2/3)

Measures of Centrality – Median (3/4):

Example 1:

Arrange all observations in order of size $n = 9$:

Order	12.3	12.6	12.6	12.9	13.1	13.2	13.4	13.5	13.6
Ranks	1	2	3	4	5	6	7	8	9

$$\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases}$$

The size $n = 9$ is odd, so the median rank is $(n + 1)/2 = 10/2 = 5$

Numerical Summaries (2/3)

Measures of Centrality – Median (3/4):

Example 1:

Arrange all observations in order of size $n = 9$:

Order	12.3	12.6	12.6	12.9	13.1	13.2	13.4	13.5	13.6
Ranks	1	2	3	4	5	6	7	8	9

$$\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases}$$

The size $n = 9$ is odd, so the median rank is $(n+1)/2 = 10/2 = 5$

The median $= x_5 = 13.1$

Numerical Summaries (2/3)

Measures of Centrality – Median (4/4):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Find the median sale of these individuals.

Numerical Summaries (2/3)

Measures of Centrality – Median (4/4):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Arrange all observations in order of size $n = 16$:

order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Numerical Summaries (2/3)

Measures of Centrality – Median (4/4):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Arrange all observations in order of size $n = 16$:

order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

$$\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases}$$

Numerical Summaries (2/3)

Measures of Centrality – Median (4/4):

Example 2:

The following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed:

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Arrange all observations in order of size $n = 16$:

order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The size $n = 16$ is even, so the median ranks are $n/2$ and $n/2 + 1$ $= 16/2$ and $16/2 + 1 = 8$ and 9

Numerical Summaries (2/3)

Measures of Centrality – Median (4/4):

Example 2:

The size $n = 16$ is even, so the median ranks are $n/2$ and $n/2 + 1$ $= 16/2$ and $16/2 + 1 = 8$ and 9

order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The median = $(x_8 + x_9)/2 = (16 + 17)/2 = 16.5$

Numerical Summaries (2/3)

Measures of Centrality – Mode (1/4):

Mode

The mode of a data set is the value that occurs most frequently. There may be no mode, or conversely, there may be multiple modes.

- If all values without repeated, then NO mode.
- If founded more than one value with the same most frequency, then here more than one mode.

Numerical Summaries (2/3)

Measures of Centrality – Mode (2/4):

Example 1:

Find the mode for the following data set

$$3, 8, 5, 6, 10, 17, 19, 20, 3, 2, 11$$

Numerical Summaries (2/3)

Measures of Centrality – Mode (2/4):

Example 1:

Find the mode for the following data set

3, 8, 5, 6, 10, 17, 19, 20, 3, 2, 11

In the data set of this example, each value occurs once except the value 3, which occurs twice. Thus, *the mode* = 3.

Numerical Summaries (2/3)

Measures of Centrality – Mode (3/4):

Example 2:

Find the mode for the following data set

$$1, 7, 19, 23, 11, 12, 1, 12, 19, 7, 11, 23$$

Numerical Summaries (2/3)

Measures of Centrality – Mode (3/4):

Example 2:

Find the mode for the following data set

$$1, 7, 19, 23, 11, 12, 1, 12, 19, 7, 11, 23$$

Note that in this data set, each value occurs twice. Thus, this data set does *not have any mode*.

Numerical Summaries (2/3)

Measures of Centrality – Mode (4/4):

Example 3:

Find the mode for the following data set

$$5,\ 7,\ 12,\ 13,\ 14,\ 21,\ 7,\ 21,\ 23,\ 26,\ 5$$

Numerical Summaries (2/3)

Measures of Centrality – Mode (4/4):

Example 3:

Find the mode for the following data set

5, 7, 12, 13, 14, 21, 7, 21, 23, 26, 5

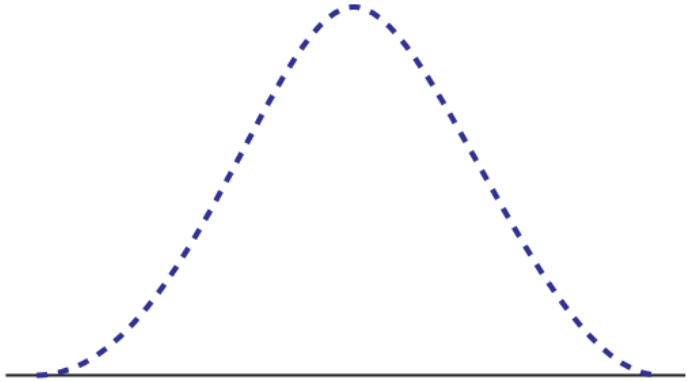
In this data set, values 5, 7, and 21 occur twice, and the rest of the values occur only once. Thus, in this example, there are *three modes* 5, 7, and 21.

Numerical Summaries (2/3)

Measures of Centrality – Shape (1/3):

Symmetric:

A data set is symmetric when the values in the data set that lie equidistant from the mean, on either side, occur with equal frequency.



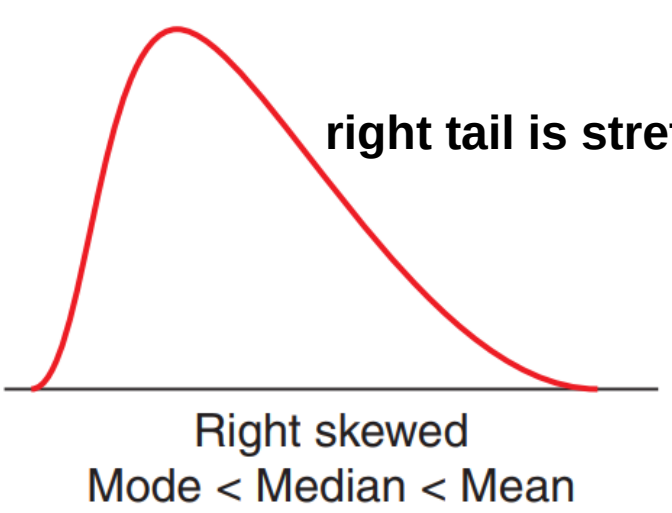
Symmetric
Mean = Median = Mode

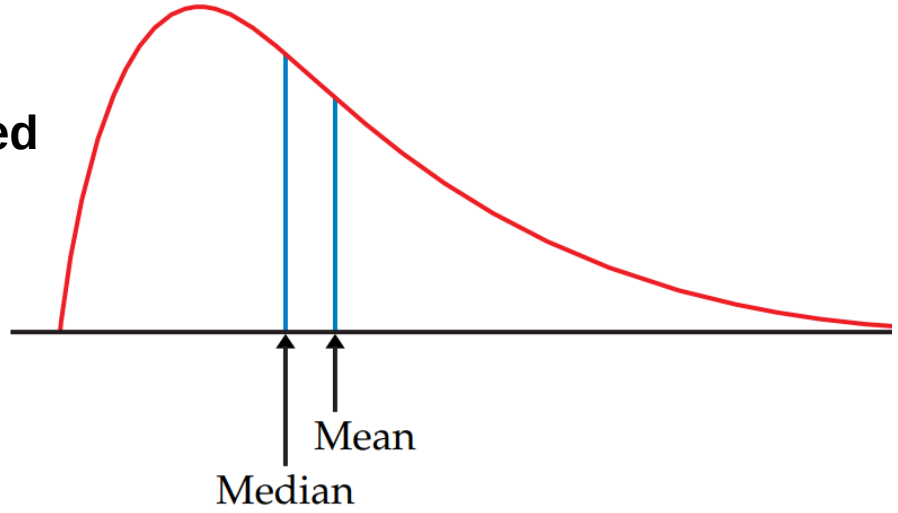
Numerical Summaries (2/3)

Measures of Centrality – Shape (2/3):

right-skewed:

A data set is right-skewed when values in the data set that are smaller than the median occur with relatively higher frequency than those values that are greater than the median. The values greater than the median are scattered to the right far from the median.



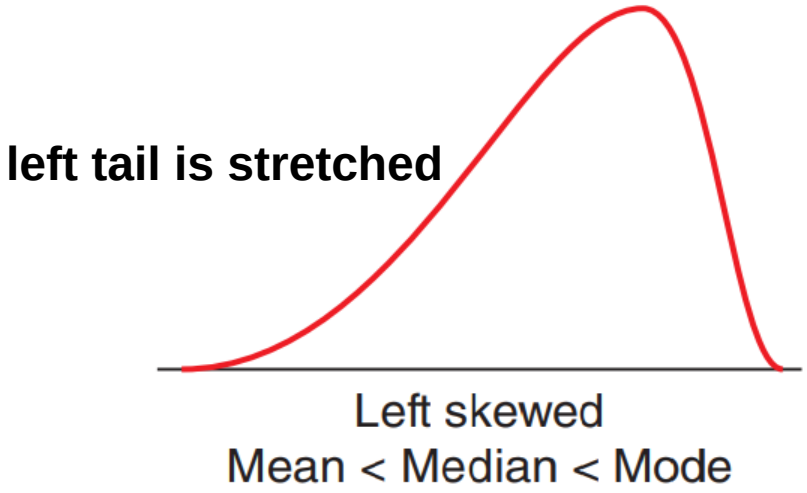


Numerical Summaries (2/3)

Measures of Centrality – Shape (3/3):

left-skewed:

A data set is left-skewed when values in the data set that are greater than the median occur with relatively higher frequency than those values that are smaller than the median. The values smaller than the median are scattered to the left far from the median.



Numerical Summaries (2/3)

Measures of Variability (Dispersion/Spread) (1/3):

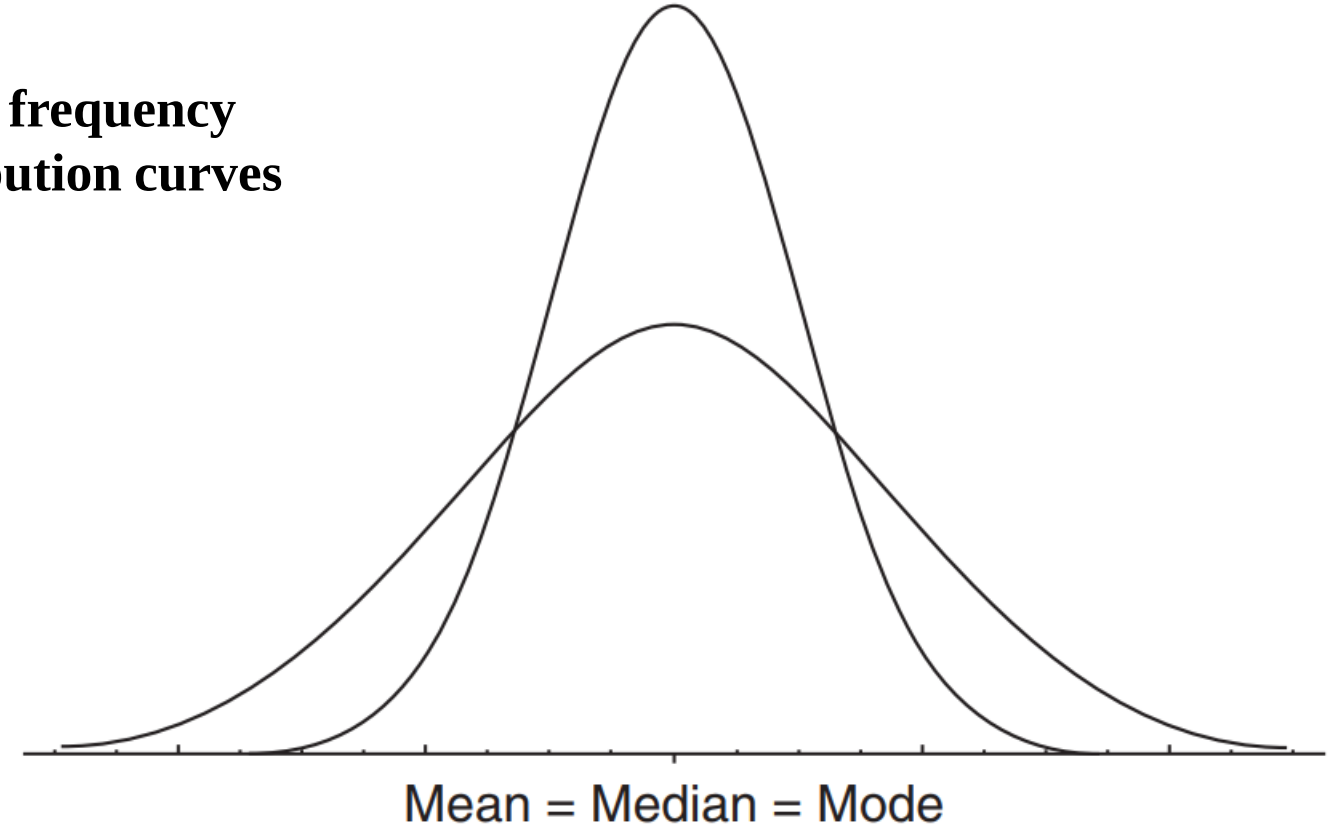
Although the sample mean is useful, it does not convey all of the information about a sample of data. Therefore, sample variability plays an important role in data analysis.

For example, a drug manufactured with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the spread or variability of incomes and drug potencies as well as their centers.

Numerical Summaries (2/3)

Measures of Variability (Dispersion/Spread) (2/3):

Two frequency distribution curves



Numerical Summaries (2/3)

Measures of Variability (Dispersion/Spread) (3/3):

We study three measures of dispersion:

1. range,
2. variance, and
3. standard deviation.

Numerical Summaries (2/3)

Measures of Variability – Range (1/2):

Range is defined as:

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Sample → Sample Range

Population → Population Range

Numerical Summaries (2/3)

Measures of Variability – Range (2/2):

Example 1:

Find the sample range for the following data set

$$1, 7, 19, 23, 11, 12, 1, 12, 19, 7, 11, 23$$

Largest value $= 23$

Smallest value $= 1$

$$\text{Sample Range} = 23 - 1 = 22$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (1/9):

Variance and Standard Deviation:

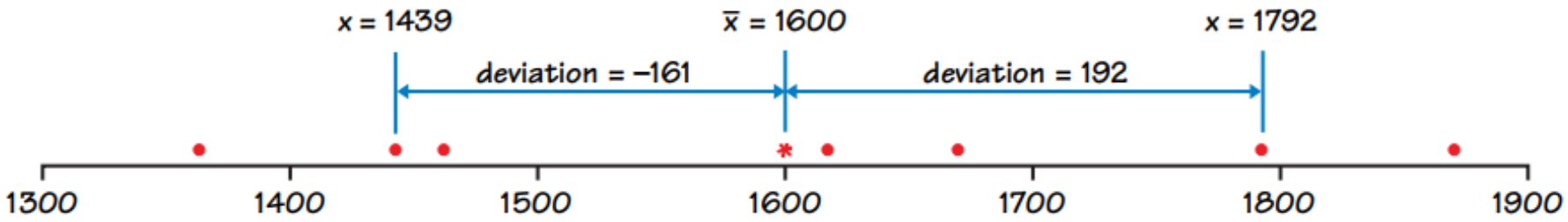
To better understand variability, we rely on more powerful indicators such as the **variance**, the **standard deviation**.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (2/9):

Variance and Standard Deviation:

The standard deviation measures spread by looking at how far the observations are from their mean.



Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

The sample variance s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the sample variance is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the sample variance is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

The ***degrees of freedom*** associated with the variance estimate

Note: the degrees of freedom depict the number of independent pieces of information available for computing variability.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If x_1, x_2, x_3, x_4, x_5 is a sample of 5 observations, and the sample mean is $\bar{x} = 90$

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
					90

$n = 5$, we have $(n - 1) = (4)$ degrees of freedom

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If x_1, x_2, x_3, x_4, x_5 is a sample of 5 observations, and the sample mean is $\bar{x} = 90$

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
Free	Free	Free	Free	Not Free	Given
					90

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If x_1, x_2, x_3, x_4, x_5 is a sample of 5 observations, and the sample mean is $\bar{x} = 90$

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
Free	Free	Free	Free	Not Free	Given
80	85	90	95		90

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If x_1, x_2, x_3, x_4, x_5 is a sample of 5 observations, and the sample mean is $\bar{x} = 90$

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
Free	Free	Free	Free	Not Free	Given
80	85	90	95		90

$$90 = \frac{80 + 85 + 90 + 95 + x_5}{5} \quad \rightarrow \quad x_5 = 100$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $80, 85, 90, 95, 100$ is a sample of 5 observations, and the sample mean is $\bar{x} = 90$. For example, suppose that we wish to compute the sample variance.

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
80	85	90	95	100	90

In general,

$$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $80, 85, 90, 95, 100$ is a sample of 5 observations, and the sample mean is $\bar{x} = 90$. For example, suppose that we wish to compute the sample variance.

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
80	85	90	95	100	90

In general,

$$\sum_{i=1}^{5}(x_i - \bar{x}) = (80 - 90) + (85 - 90) + (90 - 90) + (95 - 90) + (100 - 90) = 0$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $80, 85, 90, 95, 100$ is a sample of 5 observations, and the sample mean is $\bar{x} = 90$. For example, suppose that we wish to compute the sample variance.

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
80	85	90	95	100	90

In general,

$$\sum_{i=1}^{5}(x_i - \bar{x}) = (-10) + (-5) + (0) + (5) + (10) = 0$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $80, 85, 90, 95, 100$ is a sample of 5 observations, and the sample mean is $\bar{x} = 90$. For example, suppose that we wish to compute the sample variance.

x_1	x_2	x_3	x_4	x_5	$\bar{x}$
80	85	90	95	100	90

The computation of the variance involves,

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (-10)^2 + (-5)^2 + (0)^2 + (5)^2 + (10)^2$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

The computation of the variance involves,

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (-10)^2 + (-5)^2 + (0)^2 + (5)^2 + (10)^2$$

$\sum_{i=1}^{5}(x_i - \bar{x}) = 0$, then the computation of a sample variance does not involve n independent squared deviations from the mean $\bar{x}$. In fact, since the last value of $x - \bar{x}$ is determined by the initial $n - 1$ of them, we say that these are $n - 1$ "pieces of information" that produce s^2. Thus, there are $n - 1$ degrees of freedom rather than n degrees of freedom for computing a sample variance.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

Another way to think about $(n-1)$: In practice, the value of μ is almost never known, and so the sum of the squared deviations about the sample average $\bar{x}$ must be used instead. However, the observations x_i tend to be closer to their average, $\bar{x}$, than to the population mean, μ.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

Therefore, we use $(n-1)$ as the divisor rather than n. If we used n as the divisor in the sample variance, we would obtain a measure of variability that is on the average consistently smaller than the true population variance σ^2.

Of course, if n is large, the difference in the denominator is inconsequential.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (3/9):

Sample Variance:

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the sample variance is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}.$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (4/9):

Sample Standard Deviation:

The sample standard deviation, denoted by s, is the positive square root of s^2, that is

$$s = \sqrt{s^2}$$

Note: It is important to remember the value of s^2, and therefore of s, is always greater than zero, except when all the data values are equal, in which case $s^2 = 0$.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1: Find sample variance and sample standard deviation?

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. (meters)

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. (meters)

The sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8}$$

$$= \frac{12.6 + 12.9 + \cdots + 13.1}{8} = \frac{104}{8} = 13.0$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. (meters)

The sample mean is:

$$\bar{x} = 13 \text{ (meters)}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

$\bar{x} = 13$ (meters)

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. (meters)

The sample variance is:

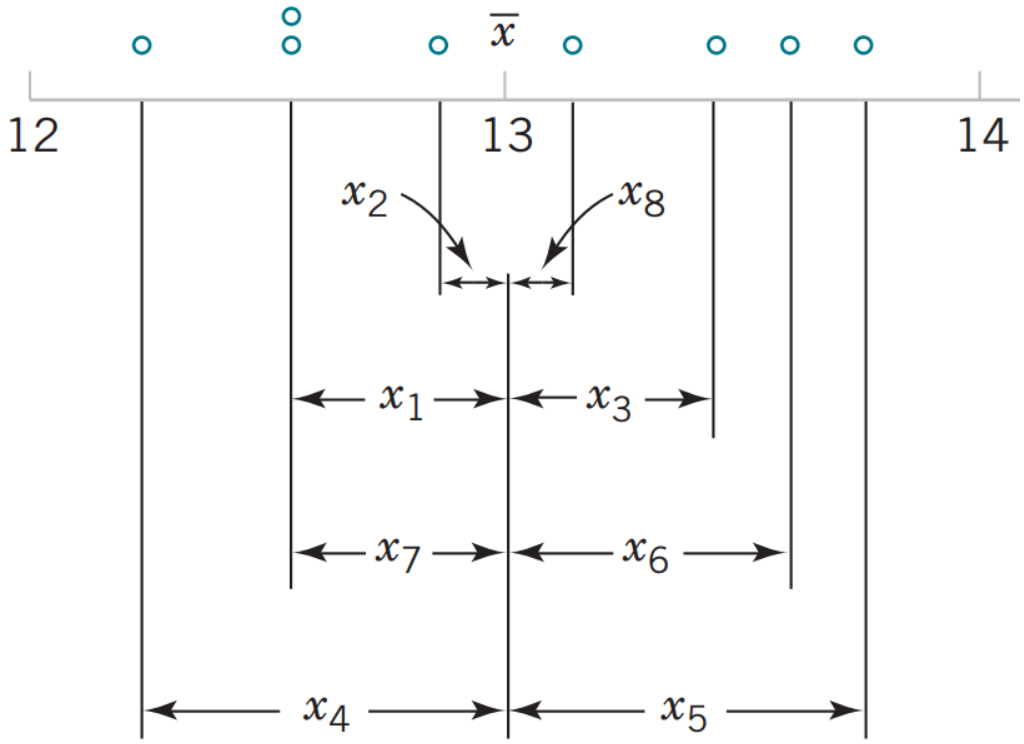
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

$\bar{x} = 13$ (meters)



$(x_i - \bar{x})$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

$\bar{x} = 13$ (meters)

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
Total	104.0	0.0	1.60

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (5/9):

Example 1:

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
Total	104.0	0.0	1.60

$$\bar{x} = 13 \text{ (meters)}$$

$$\sum_{i=1}^{8} (x_i - \bar{x})^2 = 1.60$$

the sample variance is

$$s^2 = \frac{1.60}{8-1} = \frac{1.60}{7} = 0.2286 \text{ (meters)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.478 \text{ (meters)}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (6/9):

Sample Variance:

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the sample variance is

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

The computation of s^2 requires calculation of $\bar{x}$, n subtractions, and n squaring and adding operations. If the original observations or the deviations $x_i - \bar{x}$ are not integers, the deviations $x_i - \bar{x}$ may be tedious to work with, and several decimals may have to be carried to ensure numerical accuracy.

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (6/9):

Sample Variance:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

For computational purposes, we give below the simplified form for the sample variance.

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (7/9):

Recall Example 1:

i	x_i	x_i^2
1	12.6	158.76
2	12.9	166.41
3	13.4	179.56
4	12.3	151.29
5	13.6	184.96
6	13.5	182.25
7	12.6	158.76
8	13.1	171.61
Total	**104.0**	**1,353.6**

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

the sample variance is

$$s^2 = \frac{1}{8-1}\left(1353.6 - \frac{(104)^2}{8}\right)$$

$$= 0.2286 \text{ (meters)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.478 \text{ (meters)}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (8/9):

Population Variance:

The variability in the population is defined by the population variance. When the population is finite and consists of N equally likely values, we may define the population variance as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (8/9):

Population Variance:

N is used because we assume that the population mean μ is given. Moreover, the number N is very large, the differences will be very small when use N or $N - 1$.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Numerical Summaries (2/3)

Measures of Variability – Standard Deviation (9/9):

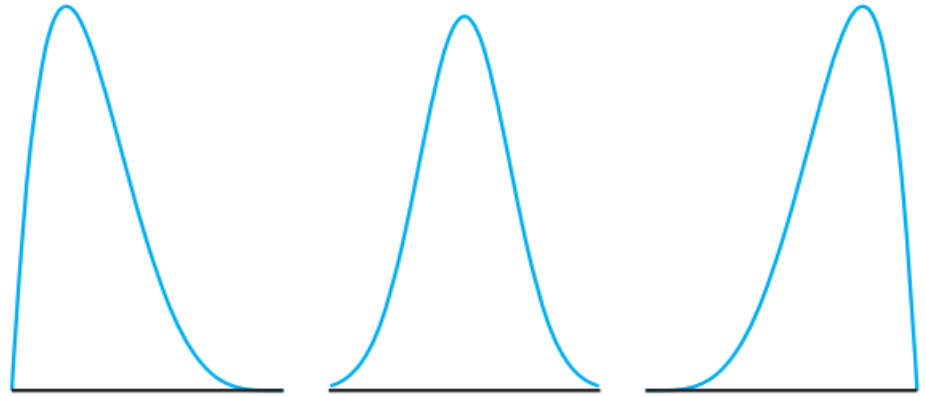
Population Standard Deviation:

The positive square root of σ^2 is the population standard deviation σ

Numerical Summaries (2/3)

Measures of Shape – Skewness (1/4):

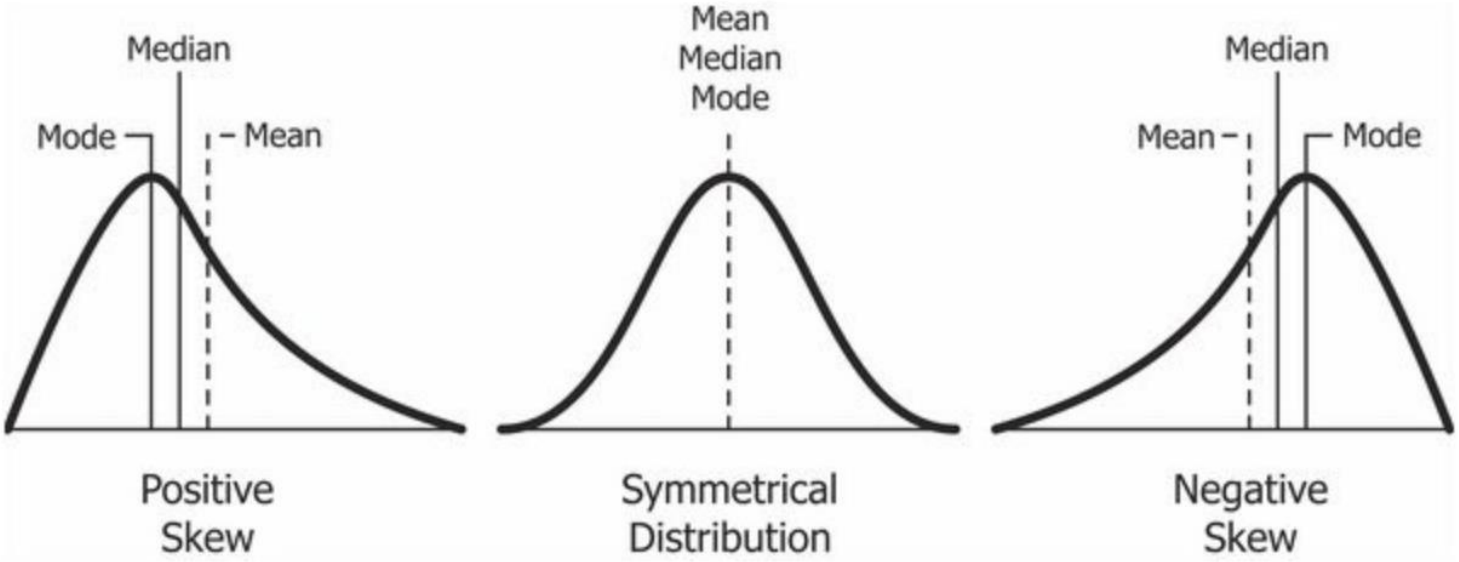
Skewness: is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.



Numerical Summaries (2/3)

Measures of Shape – Skewness (2/4):

- Right skewness is positive skewness, which means skewness > 0.
- Left skewness is negative skewness, which means skewness < 0.
- Symmetrical distribution implies zero skewness.



Numerical Summaries (2/3)

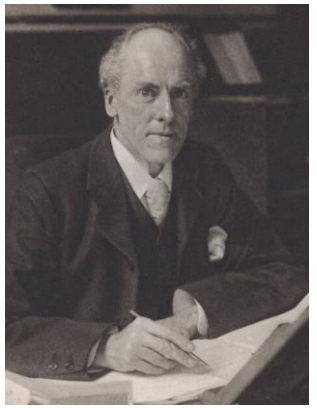
Measures of Shape – Skewness (3/4):

- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

Numerical Summaries (2/3)

Measures of Shape – Skewness (4/4):

How to calculate?



Karl Pearson, British

There are several ways to measure skewness. Pearson's first and second coefficients of skewness are two common ones.

- *Mode skewness*: Pearson's first coefficient of skewness.

$$Sk_1 = \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

- *Median skewness*: Pearson's second coefficient of skewness.

$$Sk_2 = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

Numerical Summaries (2/3)

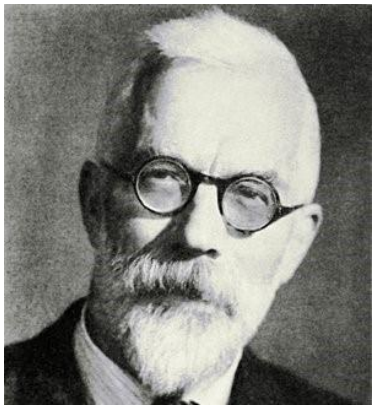
Measures of Shape – Skewness (4/4):

How to calculate?

Sample Skewness: for a sample of n values

$$g_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$

The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness.



Sir Ronald Fisher, British

Numerical Summaries (2/3)

Measures of Shape – Skewness (4/4):

How to calculate?

Sample Skewness: Many software programs actually compute the **adjusted** Fisher-Pearson coefficient of skewness. For a sample of n values

$$G_1 = \frac{n^2}{(n-1)(n-2)}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}\right)$$

Numerical Summaries (2/3)

Measures of Shape – Skewness (4/4):

How to calculate?

Sample Skewness: Many software programs actually compute the **adjusted** Fisher-Pearson coefficient of skewness. For a sample of n values

$$G_1 = \frac{n^2}{(n-1)(n-2)} \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \right)$$

g_1

Numerical Summaries (2/3)

Measures of Shape – Skewness (4/4):

How to calculate?

Sample Skewness: Alternative definition of G_1

$$G_1 = \frac{\sqrt{n(n-1)}}{(n-2)} \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \right)$$

Numerical Summaries (2/3)

Measures of Shape – Skewness (4/4):

How to calculate?

Sample Skewness:

The adjusted Fisher–Pearson standardized moment coefficient G_1 is the version found in Excel and several statistical packages including Minitab, SAS and SPSS.

Numerical Summaries (2/3)

Measures of Shape – Kurtosis (1/6):

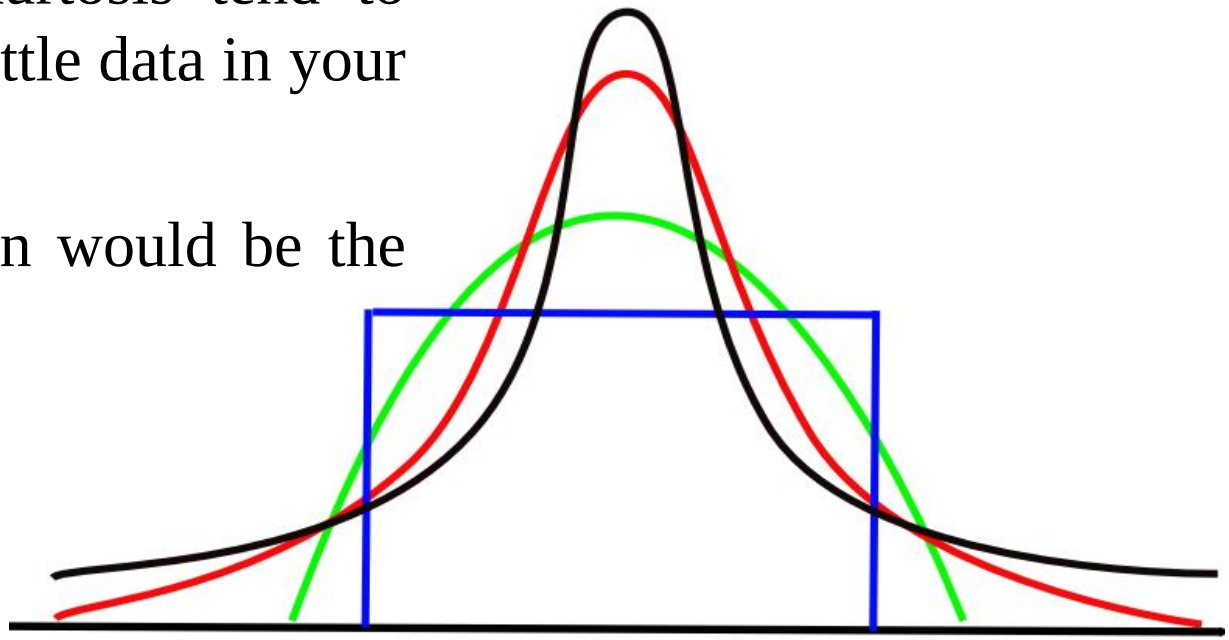
Kurtosis: is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Numerical Summaries (2/3)

Measures of Shape – Kurtosis (2/6):

- Datasets with high kurtosis tend to have heavy tails (i.e., a lot of data in your tails).
- Datasets with low kurtosis tend to have light tails (i.e., little data in your tails).
- A uniform distribution would be the extreme case.

high kurtosis
Normal kurtosis
low kurtosis
very low kurtosis



Numerical Summaries (2/3)

Measures of Shape – Kurtosis (3/6):

How to calculate?

Kurtosis, κ, can be calculated as: for a sample of n values

$$\kappa = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{s^4}$$

Greater values of κ indicates a "heavy-tailed" distribution whereas smaller κ indicates a "light-tailed" distribution.

Numerical Summaries (2/3)

Measures of Shape – Kurtosis (3/6):

How to calculate?

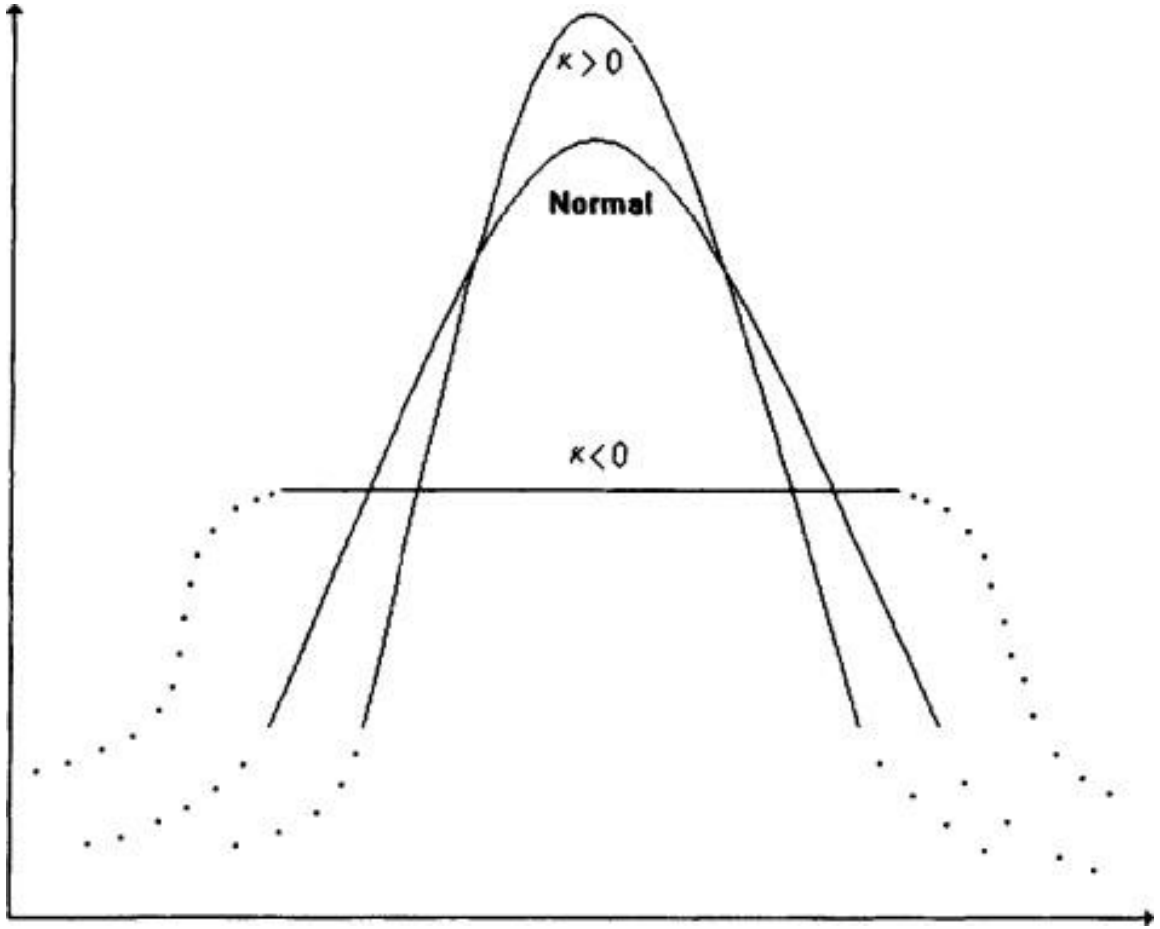
The kurtosis value for a *normal curve* often approximates **three**. For this reason, some sources use the following definition of kurtosis (often referred to as "***excess kurtosis***"): for a sample of n values

$$\kappa = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{s^4} - 3$$

This definition is used so that the normal distribution has a kurtosis of zero. In addition, with the second definition **positive** kurtosis indicates a "**heavy-tailed**" distribution and **negative** kurtosis indicates a "**light tailed**" distribution.

Numerical Summaries (2/3)

Measures of Shape – Kurtosis (4/6):

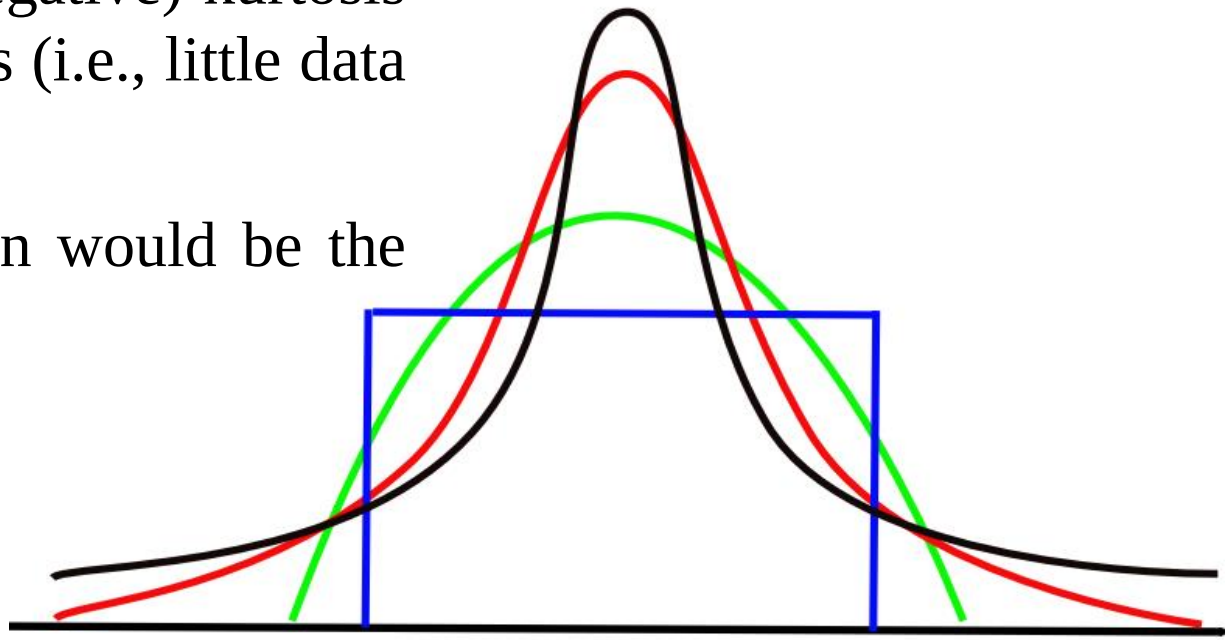


Numerical Summaries (2/3)

Measures of Shape – Kurtosis (5/6):

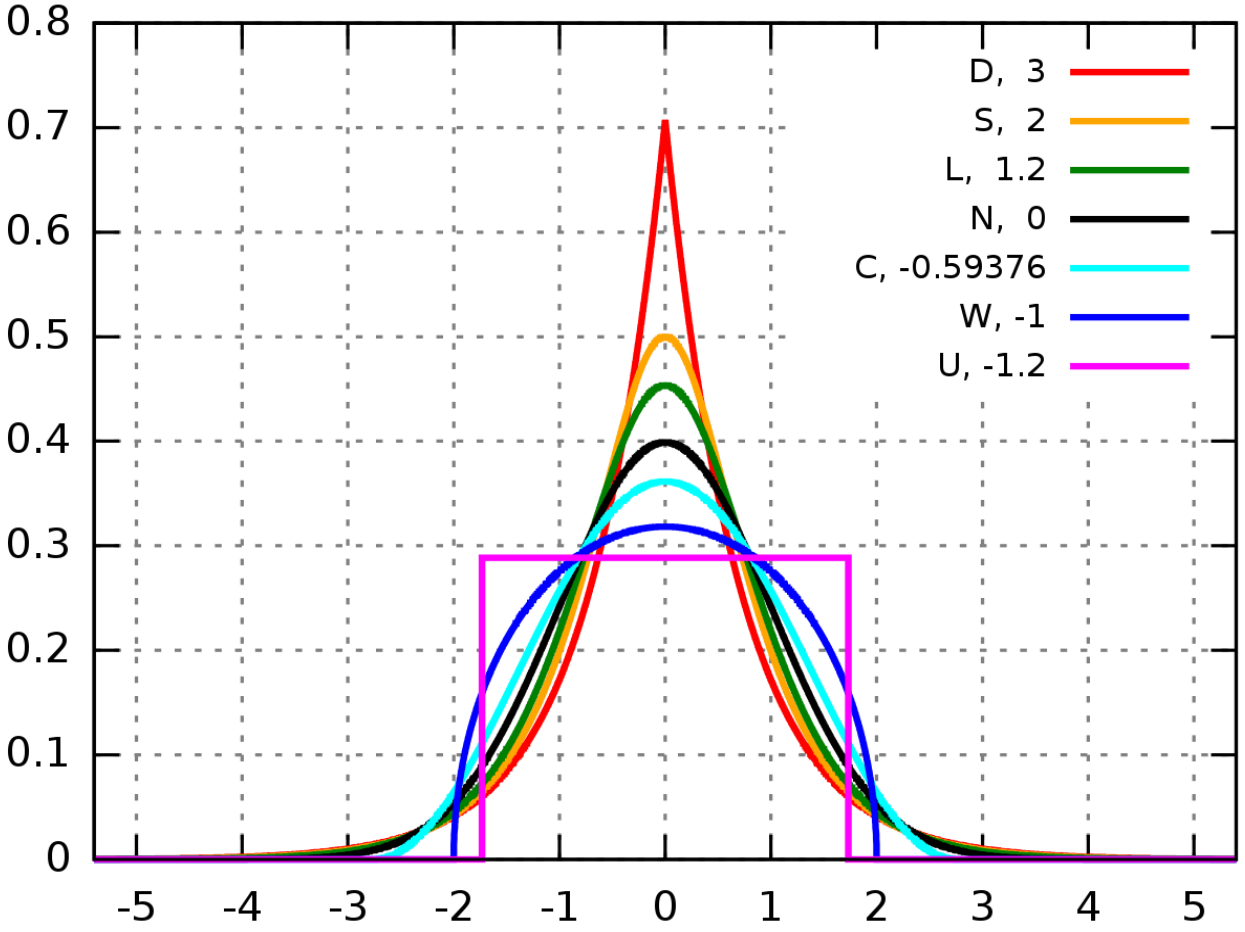
- Datasets with high (positive) kurtosis tend to have heavy tails (i.e., a lot of data in your tails).
- Datasets with low (negative) kurtosis tend to have light tails (i.e., little data in your tails).
- A uniform distribution would be the extreme case.

high (positive) kurtosis
Normal kurtosis
low (negative) kurtosis
very low (negative) kurtosis



Numerical Summaries (2/3)

Measures of Shape – Kurtosis (6/6):



Numerical Summaries (2/3)

Measures of Relative Position:

This section introduces measures of relative position that divide the data into percentages to help locate any data value in the whole data set. Commonly used measures of relative position are ***percentiles*** and ***quartiles***.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (1/4):

Percentiles: divide the data into **one hundred equal parts**; each part contains at the most 1% of the data and is numbered from 1 to 99.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	.	.	.	x_{95}	x_{96}	x_{97}	x_{98}	x_{99}	x_{100}
	1	2													99	

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (2/4):

Percentiles: divide the data into **one hundred equal parts**; each part contains at the most 1% of the data and is numbered from 1 to 99.

For example, the median of a data set is the 50th percentile, which divides the data into two equal parts so that at most 50% of the data fall below the median and at most 50% of the data fall above it. The procedure for determining the percentiles is similar to the procedure used for determining the median.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (3/4):

We compute the percentiles as follows:

Step 1. Write the data values in an ascending order and rank them from 1 to n.

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1:

(Engineers' salaries) The following data give the salaries (in thousands of dollars) of 15 engineers in a company:

62 48 52 63 85 51 95 76 72 51 69 73 58 55 54

(a) Find the 50th percentile for these data (i.e., median).

(b) Find the 70th percentile for these data.

(c) Find the percentile corresponding to the salary of $60,000.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (a) Find the 50th percentile for these data (i.e., median).

Step 1. Write the data values in an ascending order and rank them from 1 to 15.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (a) Find the 50th percentile for these data (i.e., median).

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 50th percentile = $50 \times [(15 + 1)/100] = 8$

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (a) Find the 50th percentile for these data (i.e., median).

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 50th percentile = $50 \times [(15 + 1)/100] = 8$.

The 50th percentile is 62.

That is, at most 50% of the engineers are making less than \$62,000 and at most 50% of the engineers are making more than \$62,000.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (b) Find the 70th percentile for these data.

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 70th percentile = $70 \times [(15 + 1)/100] = 11.2$.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (b) Find the 70th percentile for these data.

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 70th percentile = $70 \times [(15 + 1)/100] = 11.2$.

The 70th percentile is $\big(72 + (73 - 72)(0.2)\big) = 72.2$.

That is, at most 70% of the engineers are making less than \$72,200 and at most 30% of the engineers are making more than \$72,200.

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (c) Find the percentile corresponding to the salary of $60,000.

Now we want to find the percentile p corresponding to a given value x. This can be done by using the following formula:

$$p = \frac{\text{Number of data values} \leq x}{n+1} \times 100$$

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (c) Find the percentile corresponding to the salary of $60,000.

Now we want to find the percentile p corresponding to a given value x. This can be done by using the following formula:

$$p = \frac{\text{Number of data values} \leq x}{n+1} \times 100$$

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (c) Find the percentile corresponding to the salary of $60,000.

Now we want to find the percentile p corresponding to a given value x. This can be done by using the following formula:

$$p = \frac{\text{Number of data values} \leq x}{n+1} \times 100$$

7 values

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (c) Find the percentile corresponding to the salary of $60,000.

Now we want to find the percentile p corresponding to a given value 60. This can be done by using the following formula:

$$p = \left(\frac{7}{15+1}\right) \times 100 = 43.75 \approx 44$$

7 values

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Percentiles (4/4):

Example 1: (c) Find the percentile corresponding to the salary of \$60,000.

Now we want to find the percentile p corresponding to a given value 60. This can be done by using the following formula:

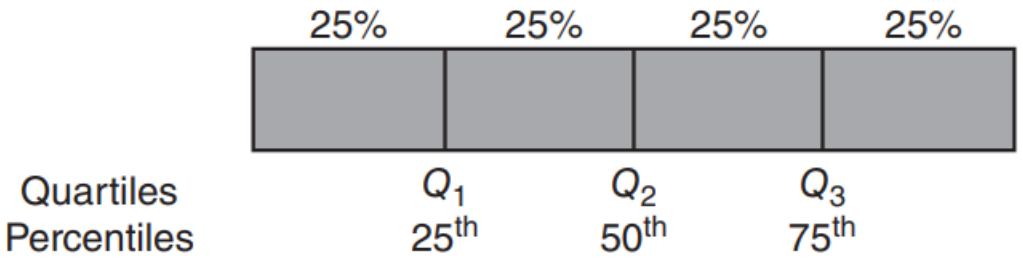
$$p = \left(\frac{7}{15+1}\right) \times 100 = 43.75 \approx 44$$

Hence, the engineer who makes a salary of \$60,000 is at the 44th percentile. In other words, at most 44% of the engineers are making less than \$60,000, or at most 56% are making more than \$60,000.

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (1/2):

Quartiles: In the percentiles, we divide the data into 100 equal parts. Some of the percentiles have special importance, such as the 25th, 50th, and 75th percentiles. These percentiles are also known as the first, second, and third quartiles (denoted by Q_1, Q_2, and Q_3), respectively. Sometimes, they are also known as the lower, middle, and upper quartiles, respectively. The second quartile is the same as the median.



Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1:

(Engineers' salaries) The following data give the salaries (in thousands of dollars) of 15 engineers in a company:

62 48 52 63 85 51 95 76 72 51 69 73 58 55 54

(a) Find the 25th percentile (1st quartile) for these data.

(b) Find the 50th percentile (2nd quartile) for these data (i.e., median).

(c) Find the 75th percentile (3rd quartile) for these data.

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (a) Find the 25th percentile for these data (i.e., Q_1).

Step 1. Write the data values in an ascending order and rank them from 1 to 15.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (a) Find the 25th percentile for these data (i.e., Q_1).

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 25th percentile

$= 25 \times [(15 + 1)/100] = 1/4(15 + 1) = 4$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (a) Find the 25th percentile for these data (i.e., Q_1).

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 25th percentile = $25 \times [(15 + 1)/100] = 4$.

The 25th percentile is 52.

That is, the 1st quartile, $Q_1 = 52$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (b) Find the 50th percentile for these data (i.e., median).

Step 1. Write the data values in an ascending order and rank them from 1 to 15.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (b) Find the 50th percentile for these data (i.e., median).

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 50th percentile

$= 50 \times [(15 + 1)/100] = 1/2(15 + 1) = 8$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (b) Find the 50th percentile for these data (i.e., median).

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 50th percentile = $50 \times [(15 + 1)/100] = 8$.

The 50th percentile is 62.

That is, the 2nd quartile, $Q_2 = 62$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (c) Find the 75th percentile for these data (i.e., Q_3).

Step 1. Write the data values in an ascending order and rank them from 1 to 15.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (c) Find the 75th percentile for these data (i.e., Q_3).

Step 2. Find the rank of the pth percentile (p = 1, 2, ... , 99), which is given by Rank of the pth percentile = $p \times [(n + 1)/100]$.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 75th percentile

$= 75 \times [(15 + 1)/100] = 3/4(15 + 1) = 12$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1: (c) Find the 75th percentile for these data (i.e., Q_3).

Step 3. Find the data value that corresponds to the rank of the pth percentile.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

The rank of the 75th percentile = $75 \times [(15 + 1)/100] = 12$.

The 75th percentile is 73.

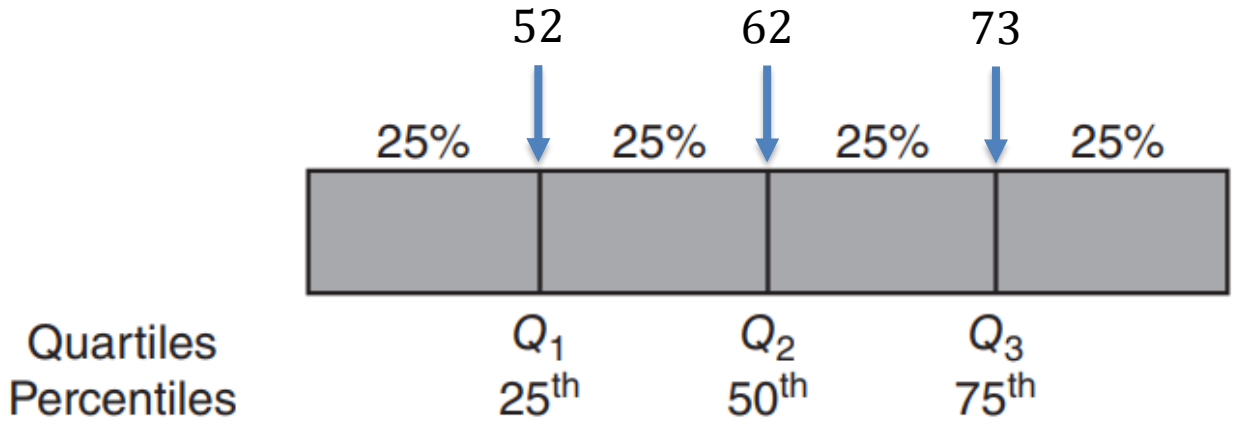
That is, the 3rd quartile, $Q_3 = 73$

Numerical Summaries (2/3)

Measures of Relative Position – Quartiles (2/2):

Example 1:

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15



Numerical Summaries (2/3)

Measures of Relative Position – IQR (1/3):

Interquartile Range (IQR): the spread between the first quartile and the third quartile, which is IQR and is defined as:

$$IQR = Q_3 - Q_1$$

Numerical Summaries (2/3)

Measures of Relative Position – IQR (2/3):

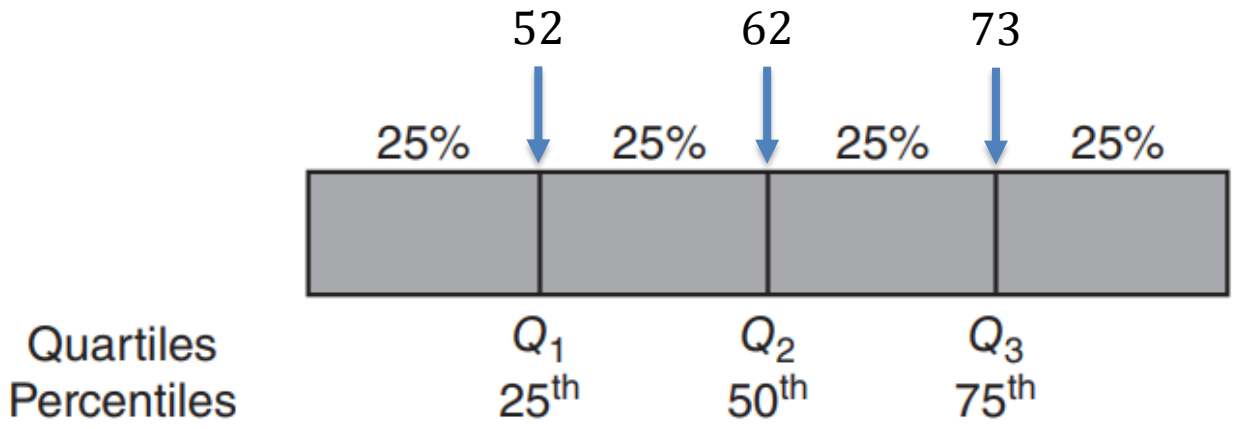
- The IQR gives an estimate of the range of the middle 50% of the population.
- The IQR is potentially a more meaningful measure of dispersion than the range as it is not affected by the extreme values that may be present in the data. By trimming 25% of the data from the bottom and 25% from the top, we eliminate the extreme values that may be present in the data set. Thus, the IQR is often used as a measure of comparison between two or more data sets on similar studies.

Numerical Summaries (2/3)

Measures of Relative Position – IQR (3/3):

Example 1:

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15



$$IQR = Q_3 - Q_1 = 73 - 52 = 11$$

Numerical Summaries (2/3)

Measures of Relative Position – cv (1/4):

Coefficient of Variation (cv): The coefficient of variation is usually denoted by cv and is defined as the ratio of the standard deviation to the mean expressed as a percentage:

$$cv = \frac{standard\ deviation}{|mean|} = \frac{s}{|\bar{x}|} \times 100\%$$

the absolute value of the mean

Numerical Summaries (2/3)

Measures of Relative Position – cv (2/4):

The coefficient of variation is a relative comparison of a standard deviation to its mean and is unitless. The cv is commonly used to compare the variability in two populations.

Numerical Summaries (2/3)

Measures of Relative Position – cv (3/4):

For example, we might want to compare the disparity of earnings for technicians who have the same employer but work in two different countries. In this case, we would compare the coefficient of variation of the two populations rather than compare the variances, which would be an invalid comparison. The population with a greater coefficient of variation, generally speaking, has more variability than the other.

Numerical Summaries (2/3)

Measures of Relative Position – cv (4/4):

Example 1:

A company uses two measuring instruments, one to measure the diameters of ball bearings and the other to measure the length of rods it manufactures. The quality control department of the company wants to find which instrument measures with more precision. To achieve this goal, a quality control engineer takes several measurements of a ball bearing by using one instrument and finds the sample average $\bar{x}$ and the standard deviation s to be 3.84 and 0.02 mm, respectively. Then, he takes several measurements of a rod by using the other instrument and finds the sample average $\bar{x}$ and the standard deviation s to be 29.5 and 0.035 cm, respectively. *Estimate the coefficient of variation from the two sets of measurements*.

Numerical Summaries (2/3)

Measures of Relative Position – cv (4/4):

Example 1:

Instrument 1: $\bar{x} = 3.84$ and $s = 0.02$

$$cv_1 = \frac{s}{|\bar{x}|} \times 100\% = \frac{0.02}{3.84} \times 100\% = 0.521\%$$

Instrument 2: $\bar{x} = 29.5$ and $s = 0.035$

$$cv_2 = \frac{s}{|\bar{x}|} \times 100\% = \frac{0.035}{29.5} \times 100\% = 0.119\%$$

The measurements of the lengths of rod are relatively less variable than of the diameters of the ball bearings. Therefore, we can say the data show that instrument 2 is more precise than instrument 1.

Numerical Summaries (3/3)

Software (1/8):

*A numerical **summary** of a distribution*:

The summary consisting of the median, the quartiles, and the smallest and largest individual observations provides a quick overall description of a distribution.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.00	9.75	15.50	15.10	20.25	23.00

Numerical Summaries (3/3)

Software (2/8):

R is a programming language and free software environment for statistical computing and graphics. Moreover, software such as Minitab, SAS, SPSS, Stata, *etc*. can be used.

Numerical Summaries (3/3)

Software (3/8):

Calculate numerical measures for the following sample data: 6, 8, 12, 9, 14, 18, 17, 23, 21, 23.

Using R

```
> data <- c(6, 8, 12, 9, 14, 18, 17, 23, 21, 23)
> summary(data)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.00	9.75	15.50	15.10	20.25	23.00

Numerical Summaries (3/3)

Software (4/8):

Calculate numerical measures for the following sample data: 6, 8, 12, 9, 14, 18, 17, 23, 21, 23.

Using R

```
> data <- c(6, 8, 12, 9, 14, 18, 17, 23, 21, 23)
> mean(data)
> Median(data)
> sd(data)
> var(data)
```

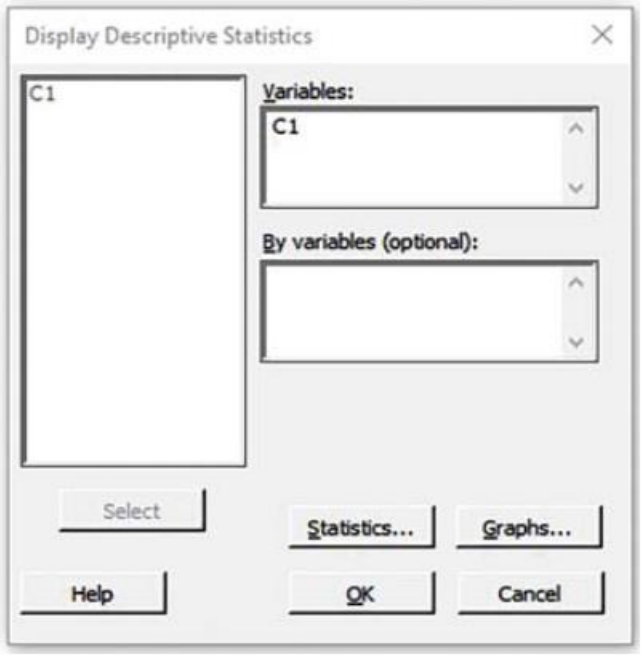
Numerical Summaries (3/3)

Software (5/8):

Minitab®

MINITAB: is a powerful statistical software everyone can use.

1. Enter the data in column C1.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. This prompts the following dialog box to appear on the screen:



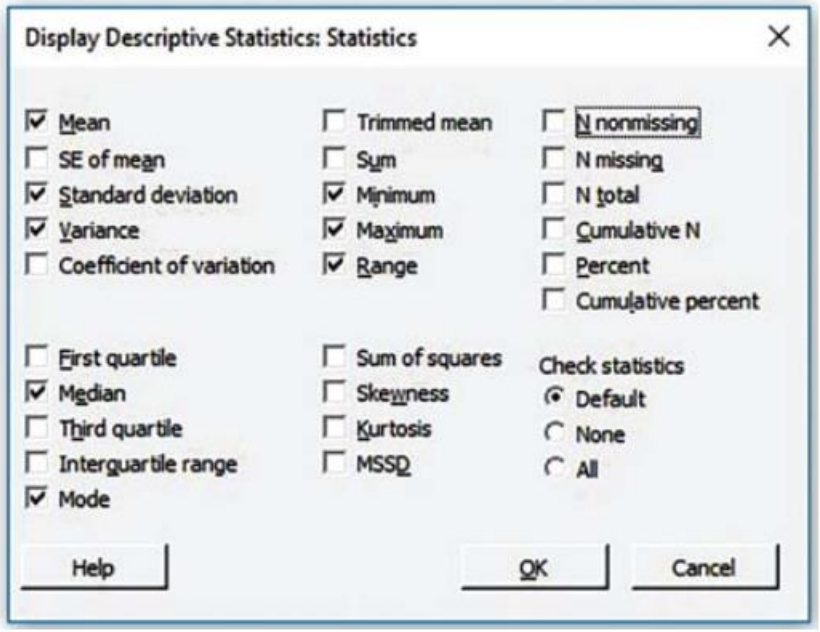
Numerical Summaries (3/3)

Software (6/8):

Minitab®

MINITAB: is a powerful statistical software everyone can use.

3. In this dialog box, enter C1 in the box under variables and click at the box **Statistics**.... Then, the following dialog box appears:



Numerical Summaries (3/3)

Software (7/8):

Minitab®

MINITAB: is a powerful statistical software everyone can use.

Statistics

Variable	Mean	StDev	Variance	Minimum	Median	Maximum	Range	Mode	N for Mode
C1	15.10	6.26	39.21	6.00	15.50	23.00	17.00	23	2

Statistics

Variable	Mean	StDev	Variance	CoefVar	Q1	Median	Q3	Range
C7	20.125	4.090	16.728	20.32	17.000	20.000	23.000	16.000

Numerical Summaries (3/3)

Software (8/8):

Using R

```
> data <- c(17,12,12,14,15,16,16,16,16,17,17,18,18,
18,19,19,20,20,20,20,20,20,20,20,21,21,21,22,22,23,
23,23,24,24,25,26,26,28,28,28)

> quantile(data)
```

0%	25%	50%	75%	100%
12	17	20	23	28

Video Lectures

All Lectures: https://www.youtube.com/playlist?list=PLxIvc-MGOs6gW9SgkmoxE5w9vQkID1_r-

Lecture #13 Part 1: https://www.youtube.com/watch?v=kie70vw-Jcs&list=PLxIvc-MGOs6gW9SgkmoxE5w9vQkID1_r-&index=15

Lecture #13 Part 2: https://www.youtube.com/watch?v=r5Vc0z67w5E&list=PLxIvc-MGOs6gW9SgkmoxE5w9vQkID1_r-&index=16

Thank You

Dr. Ahmed Hagag
ahagag@fci.bu.edu.eg